ICS-C3000 Datasta Tietoon, Autumn 2014

TENTTI 15. 12. 2014

1.

d-ulotteiset datavektorit ovat tasaisesti jakautuneita hyperpalloon, jonka säde on 1. Määritellään sisäpisteiksi ne, joiden etäisyys pallon keskipisteestä on korkeintaan $1 - \epsilon < 1$. Osoita että sisäpisteiden joukon suhteellinen tilavuus menee nolllaan kun $d \to \infty$, toisin sanoen hyvin suurissa dimensioissa melkein kaikki datapisteet ovat hyperpallon pinnalla. (Aputulos: r-säteisen d-ulotteisen hyperpallon tilavuus on $V_d(r) = C_d r^d$ missä vakio C_d ei riipu säteestä r.)

2.

On annettu otos $x(1), ..., x(n)$ suureesta, jonka tiedetään olevan normaalijakautunut

$$p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

On syytä olettaa että keskiarvo μ on lähellä nollaa. Koodataan tämä olettamus priorijakaumaan

$$p(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\mu^2}.$$

Laske Bayes-MAP-estimaatti odotusarvolle μ ja tulkitse sitä kun varianssi σ^2 vaihtelee pienestä suureen.

3.

Tarkastellaaan 1-ulotteista 3 yksikön SOM-karttaa, jonka painot ja syöte ovat skalaareja välillä [0,1]. Yksikön 1 naapuri on 2, yksikön 3 naapuri on 2, ja yksikön 2 naapurit ovat 1 ja 3. Alkutilanteessa painot ovat $m_1 = 0.5$, $m_2 = 0.25$ ja $m_3 = 0.75$. Kun uusi syöte x on valittu, etsitään ensin lähin yksikkö ja sitten sen ja sen naapureiden painoja päivitetään säännöllä

$$m_i^{uusi} = m_i + 0.5(x - m_i).$$

Valitse syöte x niin, että päivityksen jälkeen uudet painot ovat suuruusjärjestyksessä:

$$m_1^{uusi} < m_2^{uusi} < m_3^{uusi}.$$

4.

(a) Määrittele 0-1 datan kattava joukko (frequent set). Anna esimerkki pienestä 0-1-datajoukosta ja luettele sen kattavat joukot jollakin sopivalla kynnysarvolla N.

(b) Kuvaile kattavien joukkojen etsintään käytettävän tasoittaisen algoritmin toimintaperiaate.

5.

Vastaa jompaan kumpaan seuraavista esseeaiheista:

(a) hierarkinen klusterointi

(b) k:n lähimmän naapurin luokitin (k-nearest neighbour classifier).